January 13, 2025

Hon. Ona T. Wang
United States Magistrate Judge
Southern District of New York

Re: *The New York Times Company v. Microsoft Corp.* No.: 23-cv-11195: Dispute Regarding OpenAI's Responses to Document Requests

Dear Magistrate Judge Wang:

Plaintiff The New York Times Company ("The Times") seeks an order compelling OpenAI to produce the categories of documents described below.[1]

1. Documents Concerning the Market for Training Data and RAG Data

The Times seeks documents concerning "the market for [model] training data" (RFP 113) and "the market for content for retrieval-augmented generation ["RAG"] or generative search features" (RFP 114). For these RFPs, The Times also seeks documents and communications including "any discussion of the existence or non-existence of such a market."[2]

This dispute is narrow. There is no dispute the requested documents are relevant, nor could there be. *See A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1016 (9th Cir. 2001) ("Fair use, when properly applied, is limited to copying by others which does not materially impair the marketability of the work which is copied." (quoting *Harper & Row Publishers, Inc. v. Nation Enters*., 471 U.S. 539, 566-67 (1985))). Indeed, OpenAI acknowledged that its acquisition of licensed content for model training and RAG is relevant, Dkt. 147, and agreed to produce content licensing agreements and related communications, Dkt. 307 at 7, including for licensing negotiations that did not result in a final agreement.

The dispute is whether OpenAI must search for documents that explicitly address the "market" for this data. OpenAI has refused, arguing these documents will be privileged because only lawyers use the word "market," and that any review will be burdensome. Ex. 3 at 4. But OpenAI has offered no support for this assertion, and "[t]he burden of establishing the existence of an attorney-client privilege, in all of its elements, rests with the party asserting it." *Wultz v. Bank of China Ltd.*, 2013 WL 6098484, at *2 (S.D.N.Y. Nov. 20, 2013). In any event, The Times has already proposed a search term for these RFPs, which includes the word "market," and this term has yielded a reasonable number of documents (18,514). Moreover, because search terms are already subject to a target hit count, granting this request will not require OpenAI to review any additional custodial documents. The Times is simply prioritizing how to allocate the hit counts.

[1] The parties met and conferred about these disputes by video conference on November 14, November 18, January 3, and January 7. While the parties resolved some of their disputes about OpenAI's document productions, they were unable to resolve the disputes addressed in this motion. The Times's Second Set of RFPs (Requests 16-72), and OpenAI's Responses, have been filed at Dkts. 283-1 and 283-3. The Times's Third Set of RFPs (Requests 73-95), and OpenAI's Responses, have been filed at Dkts. 283-2 and 283-4. The Times's Fourth Set of RFPs (Requests 96-130) and OpenAI's Responses are attached hereto as Exhibits 1 and 2. Email correspondence between counsel is also attached hereto as Exhibit 3.

[2] RAG enables Defendants' products to integrate content from the "live" web—including The Times's website—with their large language models ("LLMs"). By using RAG, Defendants' products can generate answers to queries about current events and other information that postdates the training of the models. FAC ¶¶ 81, 108-23, 163, 179.

2. Documents Concerning OpenAI's Revenues and Profits

OpenAI should produce documents concerning revenues and profits for the at-issue models and products, including the GPT models, ChatGPT, Browse, SearchGPT, the OpenAI API, Remove Paywall, News Summarizer, Webpilot, WebGPT, and Copilot, which are relevant to at least **RFPs 67, 70, 99, and 100**[3]. *See* Dkt. 350 at 15 (OpenAI confirming these models and products are subject to discovery); Dkt. 302 at 1; Dkt. 136-1; Ex. 4 at 2. These document requests are relevant to motive, OpenAI's commercial purpose, and damages. *See* 17 U.S.C. § 504(b) (plaintiff in a copyright case is entitled to recover the infringer's profits). Yet OpenAI has agreed to provide only limited information about its revenues and profits, such as financial statements. The Times highlights three specific deficiencies.

First, First, OpenAI should be ordered to produce presentations, along with related documents and communications, regarding its revenues, profits, and valuation, as well as Microsoft's investments in OpenAI. These documents are responsive to at least RFPs 99 and 100. Yet OpenAI has limited discovery only to financial statements. Similarly, as for Microsoft's investments in OpenAI, OpenAI has limited discovery to its contractual agreements with Microsoft, excluding presentations that discuss and analyze those investments.

Such presentations and related documents are relevant and discoverable. They will provide relevant commentary and analysis lacking from financial statements and contract documents, and The Times's damages expert can use that additional information to build and support a damages model. These presentations are also relevant to the fair use analysis. For example, these presentations will reveal information about OpenAI's plans for continuing to commercialize their products and reveal the extent to which OpenAI intends to fulfill the non-profit mission upon which it was (allegedly) created. See *A&M Recs., Inc. v. Napster, Inc.*, 239 F.3d 1004, 1015 (9th Cir. 2001) ("This 'purpose and character' element also requires the district court to determine whether the allegedly infringing use is commercial or noncommercial."). These presentations will also describe and analyze trends within OpenAI's financial metrics, shedding light on which models and products have been most lucrative, and why. That information is relevant to assessing how OpenAI has benefitted from its reliance on high-quality Times content.

Relatedly, OpenAI should produce presentations about Microsoft's investments in OpenAI. The Times alleges contributory infringement because Microsoft has been, and continues to be, intimately involved in the training, development, and commercialization of Defendants' large-language models and products, including by making substantial investments in OpenAI, and The Times alleges vicarious infringement because Microsoft has profited from OpenAI's infringement. FAC ¶¶ 66, 151, 169. These presentations will provide more information than the parties' contractual agreements alone, including because such presentations will describe how Microsoft's investments have played out in practice to benefit OpenAI.

[3] **RFP 67** seeks "Documents sufficient to show (i) projected revenue streams from Defendants' Generative AI Products, (ii) past, present, and anticipated future subscriber numbers, and (iii) product-level revenue data from 2021 to present." Dkt. **RFP 70** seeks "Documents concerning referral, affiliate, advertising, and paid search revenue generated by Defendants' Generative AI Products and Services." **RFP 99** seeks "Documents concerning any internal or external presentations regarding the growth, size, earning potential, valuation, and profitability of 1) OpenAI [and] 2) Microsoft's investment in OpenAI." **RFP 100** seeks "Documents concerning any internal or external presentations showing profits and losses related to Defendants' Generative AI Products and Services."

Second, OpenAI should be ordered to produce information about its projected revenue streams and future profits, which are recoverable. *See Looney Ricks Kiss v. Bryan*, 2010 WL 5175167, at *2 (W.D. La. Dec. 7, 2010) ("[T]his Court finds as a matter of law that copyright holders can pursue future profits"). Such information is responsive to at least RFPs **67**, **99**, and **100**. In meet-and-confers, OpenAI has argued that anticipated revenue streams and future profits are entirely irrelevant to copyright damages, relying on cases where courts rejected damages models as overly "speculative." *See* Ex. 3 at 17 (email from OpenAI's counsel). But that argument is premature. The Times is entitled to discovery about OpenAI's current and anticipated revenue streams and profits, and The Times and its experts can evaluate that discovery as part of its damages model. It is improper for OpenAI to describe its anticipated revenue streams and future profits as "speculative" while withholding any discovery, especially here where The Times alleges ongoing misconduct (FAC ¶ 167).

Third, in response to RFP 70, OpenAI should be ordered to produce information about its referral, affiliate, advertising, and paid search revenue, which are also relevant to The Times's damages. *See* FAC ¶ 5 ("By providing Times content without The Times's permission or authorization, Defendants' tools undermine and damage The Times's relationship with its readers and deprive The Times of subscription, licensing, advertising, and affiliate revenue.").

3. Documents Produced to Government Authorities

In response to **RFP 106**, OpenAI should identify any domestic or foreign governmental authorities to whom OpenAI has produced documents that are relevant to the allegations and defenses in this case, and OpenAI should produce those documents.[4] *See, e.g.*, *Waldman v. Wachovia Corp.*, 2009 WL 86763, at *1-2 (S.D.N.Y. Jan. 12, 2009) (ordering production of materials provided to regulators because the "burden is slight when a defendant has already found, reviewed and organized the documents").

OpenAI has refused to even investigate the existence of a non-U.S. proceeding in which it has produced relevant documents, despite public information suggesting there may be such proceedings.[5] Accordingly, The Times seeks an order requiring OpenAI to undertake that investigation, and produce responsive and relevant documents discovered as part of that investigation. OpenAI contends that any documents produced to non-U.S. authorities would be irrelevant insofar as those authorities are not applying U.S. law. That argument does not follow. For example, if OpenAI produced documents to a foreign entity concerning OpenAI's model training practices, or provided factual information about those practices in an interrogatory-like format, that factual information would be relevant, regardless of whether the information is being applied to a different set of laws.

[4] **RFP 106** seeks "Documents submitted to any government entity, in both the U.S. and elsewhere, that are related to any of the allegations in the Complaint or the defenses thereto, including (i) copyright and intellectual property laws or rules; (ii) Defendants' business plans and monetization strategies related to generative AI; (iii) risks associated with the development of generative AI; (iv) OpenAI's non-profit status; and (v) the relationship between the Defendants." This dispute is related to another pending motion addressing The Times's **RFP 2**, which requests similar documents but is limited to United States proceedings. Dkt. 322.

[5] https://www.theguardian.com/technology/2024/jan/08/ai-tools-chatgpt-copyrighted-material-openai (discussing a submission to the House of Lords in which "OpenAI said it could not train large language models such as its GPT-4 model – the technology behind ChatGPT – without access to copyrighted work.").

Respectfully,

/s/ Ian Crosby
Ian B. Crosby
Susman Godfrey L.L.P.

cc: All Counsel of Record (via ECF)